COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13288

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13288

Machine Impostors Can Avoid Human Detection and Interrupt the Formation of Stable Conventions by Imitating Past Interactions: A Minimal Turing Test

Thomas F. Müller,[a] Levin Brinkmann,[a] James Winters,[b] Niccolò Pescetelli[c,d]

[a] *Center for Humans and Machines, Max Planck Institute for Human Development*
[b] *School of Collective Intelligence, Mohammed VI Polytechnic University*
[c] *Department of Humanities and Social Sciences, New Jersey Institute of Technology*
[d] *PSi, People Supported Technologies Ltd.*

Received 22 April 2022; received in revised form 10 February 2023; accepted 27 March 2023

Abstract

Interactions between humans and bots are increasingly common online, prompting some legislators to pass laws that require bots to disclose their identity. The Turing test is a classic thought experiment testing humans' ability to distinguish a bot impostor from a real human from exchanging text messages. In the current study, we propose a minimal Turing test that avoids natural language, thus allowing us to study the foundations of human communication. In particular, we investigate the relative roles of conventions and reciprocal interaction in determining successful communication. Participants in our task could communicate only by moving an abstract shape in a 2D space. We asked participants to categorize their online social interaction as being with a human partner or a bot impostor. The main hypotheses were that access to the interaction history of a pair would make a bot impostor more deceptive and interrupt the formation of novel conventions between the human participants. Copying their previous interactions prevents humans from successfully communicating through repeating what already worked before. By comparing bots that imitate behavior from the same or a different dyad, we find that impostors are harder to detect when they copy the participants' own partners, leading to less conventional interactions. We also show that reciprocity is beneficial for communicative success when the bot impostor prevents conventionality. We conclude that machine impostors can avoid detection and interrupt the formation of stable conventions by imitating past interactions, and that both

Correspondence should be sent to Thomas F. Müller, Center for Humans and Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: mueller@mpib-berlin.mpg.de

reciprocity and conventionality are adaptive strategies under the right circumstances. Our results provide new insights into the emergence of communication and suggest that online bots mining personal information, for example, on social media, might become indistinguishable from humans more easily.

Keywords: Turing test; Experimental semiotics; Convention; Reciprocity; Interaction history; Language evolution; Communication; Human–machine interaction

1. Introduction

The current prevalence and diversity of bots in society are unprecedented (Rahwan et al., 2019) and humans have to increasingly interact with them, especially online. Due to the nature of online environments, which often lack direct face-to-face interactions, a pressing issue concerns the ability of humans to adequately discern whom they are interacting with. This is especially problematic in instances where artificial agents fail to disclose their real identity and pretend to be human. The result can be wrong expectations or even outright deception and manipulation, as in the case of fake dating profiles (Al-Rousan et al., 2020). In response, the European Commission (2021) is discussing an obligation for AI bots to identify themselves to humans. Meanwhile, the amount of personal information that is accessible online is also growing, for instance, via public interactions with social media. In a worst-case scenario, impostor bots could learn from this information to attempt personalized frauds (Seymour & Tully, 2016), or impersonate acquaintances online to seem more trustworthy (Jagatic, Johnson, Jakobsson, & Menczer, 2007). Can access to the personal interaction history of a human make a bot impostor more deceptive? A better understanding of human strategies to distinguish humans from artificial agents could inform more targeted approaches to mediate the risk of deceptive bots.

The Turing test is the classic thought experiment to test whether a machine could pass as a human (Turing, 1950). In its modern conceptualization, normally, a human judge conducts a conversation with both the machine and another human, while trying to find out about their identity. If the machine is indistinguishable from the human, it passes the test. Notably, the formulation of this test involves natural language. Requiring human-level use of language in conversation can be seen as a particularly hard task for a machine, because what it has to display are at least two different skills: (1) the human-like ability to interact spontaneously, reciprocally, and in a relevant way even to unexpected or unusual signals (i.e., pragmatics); (2) knowledge about the conventional use of a specific language (e.g., syntax, morphology, and semantics). In this view, language is a cultural system of conventions built on human-specific abilities to communicate spontaneously by ostension and inference (Scott-Phillips, 2015; Sperber & Wilson, 1994).

The concrete progress in the application of chatbots to the Turing test mirrors these different skills: While, “first and foremost, a good grammar unit is necessary” (Pinar Saygin, Cicekli, & Akman, 2000, p. 508), recent studies have emphasized the role of pragmatics as the key challenge in passing the test (Jacquet & Baratgin, 2021). The basic argument is that mind-reading and joint co-construction are the missing abilities to construct bots that can stay relevant in a conversation with a human (Jacquet, Jamet, & Baratgin, 2021; Kopp & Krämer, 2021). What happens if we remove natural language from the Turing test? Without recourse to a conventional linguistic system that is evolved and adapted to human communicative needs, even very simple machine impostors might be convincing. This is because the threshold of what constitutes plausibly human behavior is significantly reduced, since the established communicative strategies for humans are severely limited. Consequently, this also opens up space to study how human communication can arise and stabilize in the first place. We build and expand on the insights from two different experimental approaches to human interaction, the *perceptual crossing paradigm* and *experimental semiotics*, and base our concepts of *reciprocity* and *conventionality* on them. By reciprocity, we mean interdependence between the actions of interaction partners, and we define conventions as repeated and arbitrary solutions to a coordination problem. The aim of our study is to cut natural language from a Turing test setup to gain more insight into the roles of reciprocity and conventionality in emergent communication. Our main hypotheses are that imitation of a human's past interactions (i.e., the *interaction history*) will make a bot impostor more deceptive and interrupt the successful formation of conventions.

The *perceptual crossing paradigm* is a framework investigating the identification of fellow humans through minimal interaction (for a review, see Auvray & Rohde, 2012). These experiments involve artificial, automated agents that must be distinguished from a human participant; however, they are rarely framed as Turing tests or tasks involving communication but as sensorimotor behavioral tasks instead. In the standard task, two participants can move around in a virtual 1D environment (invisible to them) and get tactile feedback whenever they touch an object in the virtual space. These objects can be either a static point in the space, the other player, or a "shadow" moving with the second player but at a fixed distance behind them. The key finding of the initial study of the paradigm (Auvray, Lenay, & Stewart, 2009) was that participants are not able to distinguish between the partner and their shadow explicitly, but the interaction between participants is more stable. Since both participants are actively looking for feedback points in the space, they oscillate back and forth when they encounter each other, leading to more interactions with one another than with the other objects in the space. This reciprocity is something the shadow object cannot deliver.

Updating the original task, Froese, Iizuka, and Ikegami (2015) demonstrated that participants are also able to explicitly become aware of the social interaction and their partner's presence when they are instructed that the task is about coordination and teamwork. Similarly, Lenay and Stewart (2012) showed this awareness also holds in instances where each object was associated with an explicit, unique sound. Other extensions of the paradigm have shown that humans can also succeed when presented with replays of authentic human behavior instead of the shadow object: Barone, Bedia, and Gomila (2020) showed this for movements drawn from a pilot human–human interaction (although only when a visual overview of the task was available). Iizuka, Ando, and Maeda (2009) found the same result even for past movements drawn from participants' partners in the interaction (although over many repeated trials, and only in a few dyads); as did Lenay and Stewart (2012), again by providing explicit feedback sounds.

In most studies using the perceptual crossing paradigm, the focus is on what we refer to as *reciprocity*, a subset of the pragmatic capabilities mentioned above. Participants must rely on social contingencies to distinguish between human partners and passive objects. Barone et al. (2020) discuss this explicitly, but it is a central feature of studies whether they measure encounter frequency (Auvray et al., 2009; Lenay, Stewart, Rohde, & Amar, 2011), turn-taking behavior (Froese et al., 2015; Iizuka et al., 2009), or pink noise patterns (Barone et al., 2020; Bedia, Aguilera, Gómez, Larrode, & Seron, 2014) to characterize the interaction between human participants. For our purposes, reciprocity is defined as the interdependence between the actions of interaction partners, that is, partner A is reactive to the behaviors of partner B, and vice versa. For true interdependence, both partners have to demonstrate reactivity to their partner. Consider a game of table tennis: Both players need to react fluently to their partner's actions to participate in the game (even if they fail to hit the ball). A non-interdependent game of table tennis would not be a game at all, like if one player left the table or stood there while not attending to the partner's serves. Likewise, a serving machine used to train players would not constitute reciprocal interaction, because it would not react to the human's returns. Similar to this example, reciprocity is an important behavioral pattern underlying mutual understanding in communication (Clark, 1996). Usually, humans will present a contribution and then iteratively react to one another until they have reached common acceptance (Clark & Wilkes-Gibbs, 1986). One prime example of such a reaction is the ability to repair misunderstandings (Clark & Schaefer, 1987), which has been shown to be a culturally universal system to resolve breakdowns in communication (Dingemanse et al., 2015).

Similar to the perceptual crossing paradigm literature, we want to investigate whether humans can leverage reciprocity to detect other humans in a virtual environment. To do so, we present them with bot impostors replaying human movement from either the participant's partner or a "foreign" participant from a different dyad. This setup is similar to one condition, respectively, from Barone et al. (2020), Iizuka et al. (2009), and Lenay and Stewart (2012), but we contrast the two conditions directly and frame the task explicitly as a Turing test. Our first research question is: Can humans successfully overcome bot impostors by interacting reciprocally in emergent communication? Given the past literature that presents reciprocity as a key behavior of "humanness" (or recognizing one another's presence in the first place), we predicted that reciprocal interaction should lead to more successful communication, no matter the condition.

Experimental semiotics is a body of literature characterized by asking participants to establish communication without access to conventional signals, such as spoken language or writing (for reviews, see, e.g., Galantucci & Garrod, 2011; Scott-Phillips & Kirby, 2010; Tamariz, 2017). The key idea is to strip conventionality away from communication to better understand the basic mechanisms it is based on. However, tasks in experimental semiotics have not usually involved bot impostors that actively interfere with the coordination process; closest to this is one example of competition between participants within a cooperative task (Santos, Matias Rodrigues, Wedekind, & Rankin, 2012).

From early on, experimental semiotics has convincingly demonstrated how humans can successfully establish novel conventions from scratch and use them for communication (e.g., Galantucci, 2005; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007). *Conventions* are defined as arbitrary solutions to a repeated coordination problem (Lewis, 1969). Consider, for instance, the convention of whether to drive on the right or left side of the road. Sticking to one side by conventional agreement massively facilitates coordination in traffic, yet for a given society, any one chosen side will function well. In dyadic interaction, conventions can emerge via explicit agreement, precedence, or perceptual salience—and, once established, can be powerful coordination devices (Clark, 1996; Schelling, 1960). Even short-term, flexible agreements, referred to as "conceptual pacts," can facilitate understanding or interfere with it, depending on precedence with an interaction partner (Metzing & Brennan, 2003). Other accounts have emphasized the role of pragmatic alignment between individuals leading to community-wide conventions (Garrod & Doherty, 1994) and likewise the role of tools to manage misalignment for converging on shared conventions (Healey, 2008).

Within the experimental semiotics literature, Scott-Phillips, Kirby, and Ritchie (2009) showed that conventions can be formed without a dedicated communication channel, through participants' virtual movements alone. Meanwhile, the form that conventions take has been shown to become more "symbolic" or "abstract" when participants have the opportunity to present feedback or repair misunderstandings (Garrod et al., 2007; Healey, Swoboda, Umata, & King, 2007). A common result in experimental semiotics is that conventions can differ wildly between participants (e.g., Healey et al., 2007; Müller, Winters, & Morin, 2019) and become more idiosyncratic over time, to the extent that they cannot be comprehended by outside observers (Caldwell & Smith, 2012; Garrod et al., 2007). In this paper, we are concerned with the emergence of these local conventions within dyads. Given specific network topologies, global coordination in a population can arise from local conventions even without individuals' knowledge (Centola & Baronchelli, 2015). Similarly, in natural language conversation, local coordination rapidly converges into global conventions in connected groups, but not in isolated pairs or disconnected groups (Garrod & Doherty, 1994).

In the perceptual crossing paradigm, local conventions that differ between dyads are not always a possibility, for instance because of constantly changing interaction partners (Barone et al., 2020). This ensures that only reciprocity is causing successful interaction. An in-depth study of local conventions within the perceptual crossing paradigm is the one by Iizuka, Marocco, Ando, and Maeda (2013), whose dyads were able to communicate a simple visual stimulus through repeated interaction. However, even in this case, a major difference to experimental semiotics is that interaction is maximally limited, such that strategies other than reciprocal oscillation are not possible. We open up space for diverse conventional strategies by presenting participants with a more open visual environment and constant interaction partners. Thus, our second research question is: Will humans successfully use conventionality to overcome bot impostors, where possible? This relates to our prediction that conventional interactions, that is, repeating past successful behaviors, would aid communicative success normally, but become a liability under the condition that bot impostors imitate a participant's partner.

The fact that conventions can become arbitrary to the point that they are unintelligible to outsiders already touches on the powerful effects of having access to a shared *interaction history* (i.e., all behaviors produced in mutual interaction). Clark (1996) notes this as one of the central features of the common ground on which humans build their communication. It has long been established that as conversation in natural language unfolds, humans leverage their interaction history to become more efficient in their communication (e.g., Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964). More recently, studies in experimental semiotics have shown that the interaction history between participants can play a critical role in establishing conventions and forming linguistic structure (Nölle, Staib, Fusaroli, & Tylén, 2018; Silvey, Kirby, & Smith, 2015; Tinits, Nölle, & Hartmann, 2017; Winters, Kirby, & Smith, 2015, 2018). With this in mind, it is reasonable to assume that accessing the interaction history might be beneficial for a bot impostor within the setting of a Turing test. Similar to how humans differentiate in conversation to avoid confusion between past and current referents (Brennan & Clark, 1996; Yoon, Benjamin, & Brown-Schmidt, 2016), bot impostors might be able to deceive humans about their identity by imitating previous interactions. This last point motivates our third research question: Does access to a pair's past interactions make an imitative bot impostor harder to detect? We predicted that this would be the case because the tendency of humans to leverage their shared interaction history will become a disadvantage for participants in this condition.

In summary, we are investigating: (i) whether access to past interactions will make a bot impostor harder to detect; (ii) whether humans will use conventionality to successfully overcome and detect bot impostors; and (iii) whether humans will use reciprocity to successfully overcome and detect bot impostors. We address these questions in a minimal Turing test experiment stripped of natural language. Participants tried to categorize their interactions as being with their human partner or a bot impostor. Crucially, we manipulated bot behavior between participant pairs such that bots either copied the movements from a foreign pair or copied the movements of a participant's human partner. In doing so, we experimentally vary the ability of bots to access participants' past interactions and use this to test its influence on performance. Specifically, we predict that it is harder for participants to succeed when bots have access to the interaction history. We make this prediction because participants should have a harder time building on past successes. Furthermore, we predict that conventionality is a less successful strategy in this condition, since the bot impostor impedes the reuse of conventions as a means for detection by sampling from previous interactions. Conversely, we expect conventions to be a successful approach in the other condition, as bots do not sample from the interaction history. Lastly, we expect reciprocal interaction to be a successful approach no matter the condition, since it does not depend on interaction history and can be dynamically negotiated to establish communication.

2. Method

2.1. Procedure

Participants ($n = 200$) were recruited online via Prolific and assigned in pairs to a random experimental condition. They could only take part if they did not participate in any of the pilot runs of the experiment and if they spoke English. A single experiment lasted for, on average, 45 min, for which participants were paid £5.13, plus an additional bonus of £0.02 for each time they correctly identified their partner as a human or bot and for each time their partner

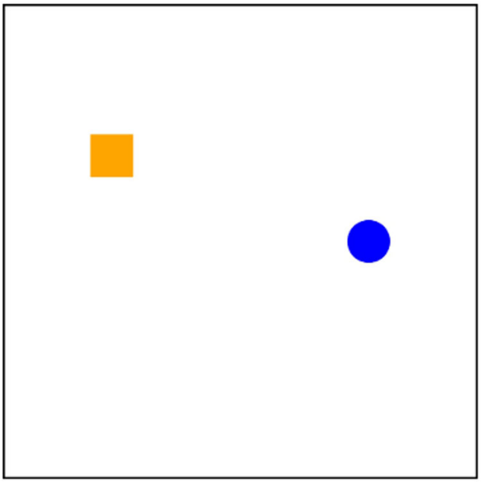

Fig. 1. The 2D space in which participants communicated. Any participant always moved the orange square, while their partner moved the blue circle. Two example interactions from a successful pair can be found in the Supplementary Material.

guessed correctly. This incentive was designed to encourage participants to cooperate and focus on pair performance rather than just their individual success. Before we started testing any participants, we received approval from the Institutional Review Board of the Max Planck Institute for Human Development. The experiment itself was designed and implemented in *Empirica* (Almaatouq et al., 2021).

Participants gave their informed consent and read the instructions before starting the experiment. They then had to answer three basic questions about the task to proceed (to ensure they had understood the instructions). The instructions explained that the experiment's goal was to find out how well humans could identify bot impostors in a virtual environment. Participants were tasked with identifying whether their current partner in a trial was human or a bot, and trying to demonstrate their own human identity to their partner. To do so, they could move around horizontally and vertically in a 2D space (Fig. 1) with their shape. No other communication or interaction between participants was possible; similar to Scott-Phillips et al. (2009), they had to rely on their virtual movements alone, without a dedicated communication channel.

At the start of the experiment, participants played three training trials, in which no partner was present and they could move around with their shape in the empty space. They were instructed to try the movement and think about how they would approach the task. These training trials served to familiarize participants with the experiment and to seed the bot's first trials in the partner impostor condition. After the training, every pair of participants played 48 trials of the task, half of which were human–human trials, while the other half were human–bot trials (no matter the experimental condition). These 48 trials were organized into six blocks of eight trials, where in every block four human–human trials and four human–bot trials were administered randomly.

In every trial, participants had 20 s to move around in the space and interact with their partner. These 20 s were preceded and followed by periods of 3 s, during which the participants could not move and saw either "Get ready" or "Round over!" on their screen. This was to give them some time to pause, and more importantly to buffer potential lag. Participants (and bots) always started with their shape in the central position within the 2D space in every trial and moved from there. They did so by using the arrow keys on their keyboard. The movement was restricted to the horizontal and vertical direction only, and was implemented such that shapes always moved the distance of their full length at a time (as if participants were on an invisible grid). Consequently, shapes always moved at the same speed, could only overlap fully (not partially), and the square and the blue circle (on top) were both visible when that happened. Every participant in the experiment used the orange square shape when they played, and always saw their partner (human or bot) as the blue circle. As such, the shapes were kept equal for all participants, even within a pair. To guard against internet lag distorting communication, we implemented two more features: First, the game moved at five frames per second, meaning that any delay smaller than 200 ms was not visible to participants and a delay smaller than 400 ms meant the screens only differed by one frame. Second, the movements of human and bot partners were both sent through the partner's client, meaning the bot partner did not differ from the human partner in lag and could not be identified that way.

After each trial, participants had 10 s to answer the two following questions: "What was the identity of your partner last round?" ("Human" or "Bot"); and "How confident are you in your choice above?" ("very confident," "rather confident," "not confident," "not at all confident"). During this period, they could change their mind on their answers, but after the time limit, the answer was submitted. After this stage had passed, participants received feedback about their and their partner's performance. For 5 s, they saw information about the correct answer for the last trial ("Human" or "Bot"), their own choice ("Human," "Bot," or "Nothing," followed with "correct" or "incorrect"), and their human partner's choice (same as last). The entire experiment ran synchronously, so participants always were at the same stage at the same time. This was true even in the training and bot rounds, meaning that during a human–human round, both players were in the trial together, but in a bot round, they had separate encounters with their bot partners (hence, there were two separate trials running synchronously).

In a between-participants design, our two conditions differed in the bot that participants encountered, but were identical otherwise (Fig. 2). In the *partner impostor condition*, bots presented faithful replays of the behavior of participants' own human partners. Bots chose a random trial from the last four human–human trials and completely copied the partner's behavior to show to the participant; however, they never chose the last trial and never chose the same trial twice in a row. If there were not four human–human trials yet, the bot would access and copy the partner's training trials. In the *foreign impostor condition*, bots presented faithful replays of human participants' behaviors from a previous experiment in the partner impostor condition. This was implemented such that the bot always replayed the same human's behavior, and in the correct order of human–human trials in the original experiment. This allowed the bot to capture the development over time that occurred in the original experiment. Bots within a pair were also parallel: One participant encountered the behavior of a previous participant as the bot and their partner saw the behavior of the original source's

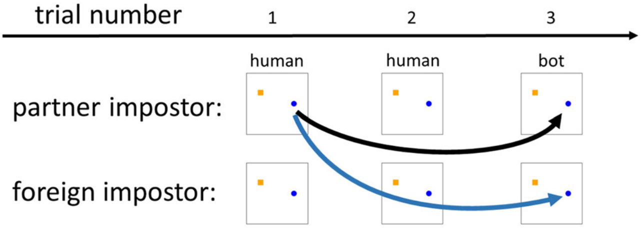

Fig. 2. Visualization of the experimental manipulation. Bots in the partner impostor condition drew from the previous behavior of a participant's human partner, while bots in the foreign impostor condition copied previous participants from a completed partner impostor condition.

partner. Participants in both conditions were aware that they would always encounter the same bot or human during the experiment.

Our aim for the manipulation in bot behavior was to vary only the bots' access to a pair's interaction history, such that we can investigate its consequences for accuracy, conventionality, and reciprocity in participants' communication. In general, pairs of participants were assigned to their condition randomly, with the exception that the foreign impostor condition could only happen whenever there was at least one "unused" partner impostor condition as a seed for the bot behavior, as each one was only used once. Bot impostors in both conditions were similar in that they presented authentic human behavior, pulled from trials that only involved humans. However, they differed in how they related to the participants' own past interactions: While the foreign impostor presented behavior from a completely unrelated dyad, the partner impostor accessed the personal interaction history. Even the concrete bot behaviors were mostly parallel between the conditions (since one was pulled from the other). The only exception to this is that the order in the foreign impostor condition resembled the true sequence of what had happened within the source pair, whereas the partner impostor condition presented randomly chosen trials of the most recent interactions. This resulted in a more stochastic representation of the original order and with not all trials replayed, which was necessary not to give the partner impostor away.

After the experimental trials had ended, the participants answered several short questions regarding their personal strategy, what they suspected the bot was doing and what they expected it to do, and whether they experienced technical difficulties. Finally, participants saw a feedback screen detailing their earned bonus payment and a summary on how they performed. This was followed by a short summary of the goals of the experiment and an explanation of what the bots were doing in the task.

2.2. *Measures*

The following measures were preregistered in advance and used in our analyses (conducted in R version 4.1.2; R Core Team, 2021). Two more descriptive measures (confidence and behavior indices) were preregistered and are reported in the Supplementary Material, along with their results.

2.2.1. Accuracy

The key measure on how accurate participants were was simply their answer to the question about the partner's identity after each round, coded as correct or wrong. Descriptively, this was complemented by a comparison to chance performance, and signal detection measures (*psycho* package; Makowski, 2018).

2.2.2. Behaviors

In general, we analyzed participants' behaviors systematically in two ways: by the set of positions in the space per trial, and by the set of directions chosen per trial, for every respective time step. The former can be represented by the x- and y-coordinates of the space (e.g., time step 1: $x = 220$, $y = 220$; time step 2: $x = 240$, $y = 220$), and the latter by a string of letters each denoting one of the five possible actions (the four directions and "stop"; e.g., "R" for "right" and "L" for "left" results in "RL" for the sequence in the first example and back).

2.2.3. Conventionality

Conventionality was assessed by comparing the similarity of behaviors over time for a given participant. In principle, more conventional behaviors should be characterized by higher stability within the game, and this stability should increase as a convention becomes more entrenched over time. Conventions can develop arbitrarily both about the set of directions and the set of positions per trial in our task, leading us to include measures for both. For the set of directions, we made use of the Levenshtein distance between the sequences. This was the number of insertions, deletions, or replacements needed to make the sequences equal (since they always had the same length of 100 time steps, no further adjustment was needed). For the set of positions, we compared the frequency distributions by using the Earth mover's distance (package *emdist*; Urbanek & Rubner, 2012). This was the number of positional changes needed to make the distributions identical in space, weighted by the Manhattan distance (again, there were always 100 positions recorded for every trial). Both measures were computed as the average distance to a participant's own behaviors within the same trial block (to smoothen random fluctuation from one trial to another).

2.2.4. Reciprocity

Quantifying true interdependence is hard because of the question of causality. Our approach measured the predictability of behavior shown by participants' partners, given a participant's own behavior in the current trial. To do so, we used the transfer entropy (*RTransferEntropy* package; Behrendt, Dimpfl, Peter, & Zimmermann, 2019), quantifying the information flow from one participant to another. We chose this measure because: (i) It is directional, allowing us to measure both the impact player A has on player B and the impact player B has on player A. (ii) It is time-delayed, which is ideal because reactions cannot happen immediately. (iii) It does not rely on the impacted player's behaviors being identical to the actions by the partner, but only on the reactions being consistent. Since the reactivity to one another is by its nature about the movements, not positions, we only focused on the set of directions (again including "stop") here. We decided a lag of two timesteps (about 400 ms) between sequences was ideal for the measure, since it produced good variance and reflects a reasonable reaction time, given that interactions were online with some connection lag. The result of the measure was a single number per participant per trial in the experiment, indicating the information flow from them to their partner. This was the only direction that made sense in the context of our study: Participants can try to decide about their partner's identity based on how much their partner reacted to their behaviors, but not based on their own behaviors.

2.3. Predictions

We preregistered three research questions and the corresponding predictions on the Open Science Framework in advance (https://doi.org/10.17605/OSF.IO/TXGH7). These were:

1. Does access to a pair's past interactions make an imitative bot impostor harder to detect?
2. Will humans successfully develop conventionality to overcome bot impostors, where possible?
3. Can humans successfully overcome bot impostors by interacting reciprocally in emergent communication?

Research question 1 concerns our measure of accuracy and the central manipulation of the experiment. We predicted that pairs in the foreign impostor condition would outperform pairs in the partner impostor condition, because they can rely on conventional in addition to reciprocal interactions (i.e., leverage their interaction history). Although it is possible, note that these interactions do not necessarily have to take the form of explicit strategies that participants are aware of: Like in the literature regarding the perceptual crossing paradigm, the degree of awareness can vary depending on the study design and we are not making any specific claims about it here. Essentially, we are linking emergent behavioral patterns to success in the task.

Research question 2 concerns our measures of conventionality and accuracy. We believed that higher conventionality would be beneficial in the foreign impostor condition, but detrimental in the partner impostor condition, because the bot would start to imitate participants' conventions as the experiment goes on. Our predictions were that (1) more conventional behavior would lead to better performance in the foreign impostor condition, but worse performance in the partner impostor condition, and (2) pairs above chance performance in the partner impostor condition would interact less conventionally than their counterpart in the foreign impostor condition. We made this second prediction because reusing successful behaviors is punished by the imitative bot in the partner impostor condition, preventing the formation of conventions.

Research question 3 concerns our measures of reciprocity and accuracy. In short, we believed that reciprocity would enable success in both experimental conditions, but be relatively more important for success in the partner impostor condition, because pairs in the foreign impostor condition should also be able to rely on conventionality (question 2). Consequently, we predicted that: (1) Overall, more reciprocal behavior would lead to more successful guesses; and (2) pairs above chance performance in the partner impostor condition would interact more reciprocally than their counterpart in the foreign impostor condition.

Additionally, we believed that the respective adaptive strategies of reciprocity and conventionality should also show an increase over time in the relevant conditions, along with an increase in accuracy as participants learn how to solve the task.

3. Results

3.1. Preprocessing

Our full data and scripts for the analyses can be accessed here: https://doi.org/10.17605/OSF.IO/Z4VTJ. As preregistered, we excluded all data from participants who did not finish the experiment because of our "failsafe" (more than three trials missing due to them leaving the experiment or losing the online connection; this concerned 16 pairs in total), and recruited participants until we met our preregistered sample size of 100 pairs. Likely because of this approach, we did not have to exclude any data due to lag spikes (over 3000 ms for at least 10 s, per the preregistration). On the trial level, however, seven trials (out of 9600) were lost due to leaving participants or lost connections.

3.2. Qualitative self-reported strategies

Participants answered to our question "Which strategies did you use to solve the task, if any?" after the game. We summarize the preferred strategies qualitatively in Chapter S4 and Table S1. References to conventional strategies commonly stated that they used "repetitions" or "similar patterns." Examples of patterns included moving to the corners of the space, circular patterns, L-shapes, and more. References to reciprocal strategies commonly mentioned "following each other" or "copying each other."

3.3. Descriptive statistics

3.3.1. Accuracy

Overall, mean performance in the task was $M = 35.4$ correct answers out of 48 trials per participant (73.7% accuracy, $SD = 8.7$; see Fig. 3), with chance level being 24 correct answers (50%). Participants performed better in the foreign impostor condition ($M = 42.0$, 87.4% correct, $SD = 6.0$) than in the partner impostor condition ($M = 28.8$, 60.0% correct, $SD = 5.2$), and they improved more over time (Fig. 3). Forming 95% confidence intervals around the 50% mark, 48 pairs in the foreign impostor condition and 22 pairs in the partner impostor condition performed above chance. Performance was better in the 24 trials participants played with the human partner ($M = 19.0$, 79.1%, $SD = 4.2$) than in the trials with the bot partner ($M = 16.4$, 68.3%, $SD = 5.4$). The mean sensitivity (d') for accurately identifying the partner (with "positive" being a human partner) was 1.5 overall ($SD = 1.3$), further indicating that participants made more correct identifications than incorrect ones. This was again higher in the foreign impostor condition ($M = 2.5$, $SD = 1.0$) than in the partner impostor condition ($M = 0.5$, $SD = 0.6$).

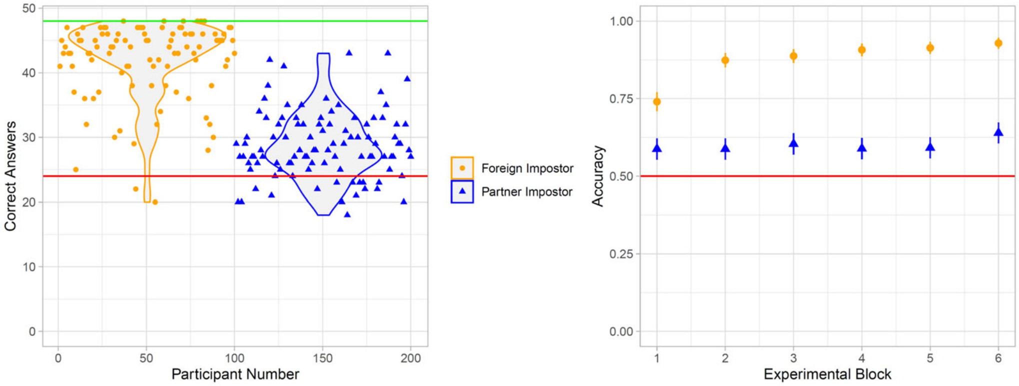

Fig. 3. Left: Distribution of performance by participants. The middle line indicates chance performance, and the top line is the performance ceiling. Right: Development of performance over experimental blocks. The line indicates chance performance; error bars are 95% confidence intervals. The foreign impostor condition started close to the partner impostor condition (in fact the first few trials were equally at chance), but quickly improved almost toward performance at ceiling, while the partner impostor condition average stayed close to chance.

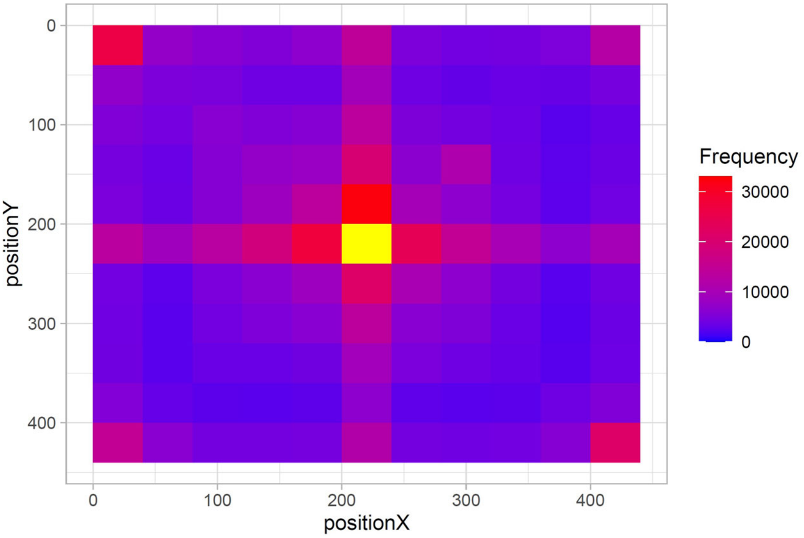

Fig. 4. Overall distribution of participants' positions in the space. Each cell corresponds to one position in the 2D space. Each time step in the game contributed one observation for both human participants' current positions (i.e., 100 observations each per trial). The center is indicated in yellow since its associated frequency value (130,067) is completely off scale (unsurprisingly, given that participants had to start there).

3.3.2. *Frequency of positions*

We computed the overall frequency distribution of positions in the space (Fig. 4; most relevant for the Earth Mover's distance as a measure for conventionality). Participants most frequently moved straight horizontally or vertically from the central starting position, and

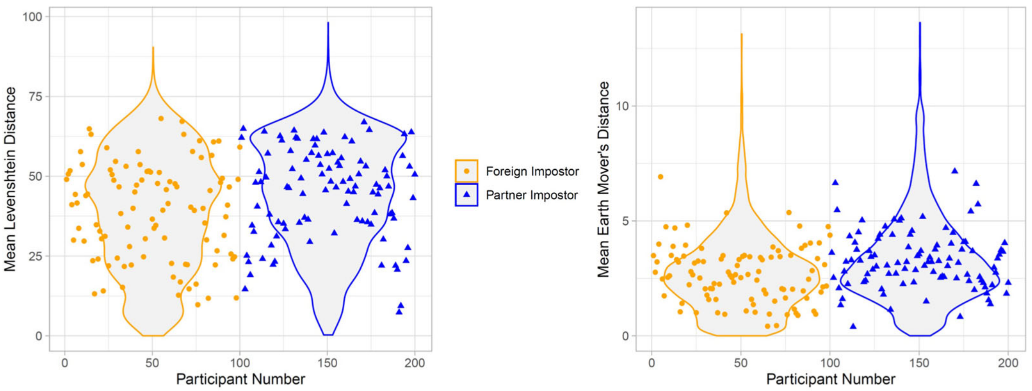

Fig. 5. Mean conventionality by participant. Left: Mean Levenshtein distance. Right: Mean Earth Mover's distance.

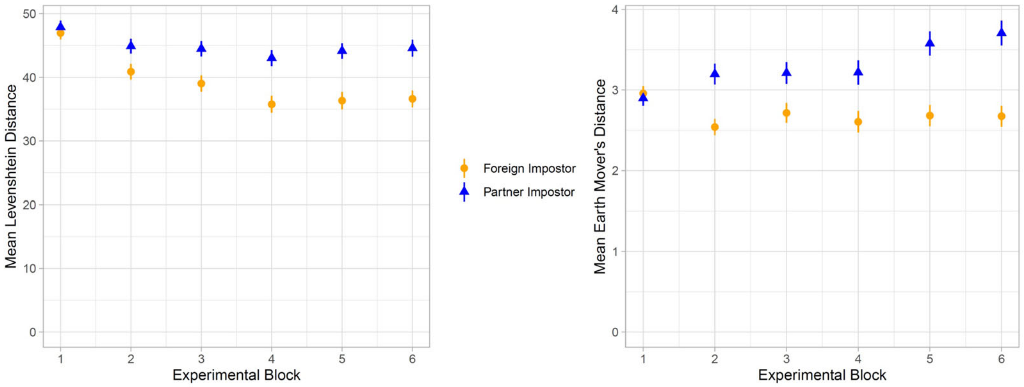

Fig. 6. Development of conventionality over experimental blocks. Error bars show the 95% confidence intervals. Left: Mean Levenshtein distance. Right: Mean Earth Mover's distance.

moved around less in peripheral than in more central positions (unsurprisingly, given the starting position). However, the exception to this pattern was that they much preferred moving along the borders of the space, and massively favored the four corners.

3.3.3. Conventionality

The average Levenshtein distance was slightly lower (meaning higher conventionality; Fig. 5) in the foreign impostor condition ($M = 39.3$, $SD = 18.4$) than in the partner impostor condition ($M = 44.8$, $SD = 17.2$), and so was the average Earth Mover's distance ($M = 2.7$, $SD = 1.7$; vs. $M = 3.3$, $SD = 2.0$; Fig. 5). Over the experimental blocks, we found the expected pattern for the Earth Mover's distance, which was increasing in the partner impostor condition (Fig. 6). In the foreign impostor condition, it decreased from block 1 to block 2, but stagnated after that. Unexpectedly, the Levenshtein distance decreased in the partner impostor condition along with the foreign impostor condition (but less so).

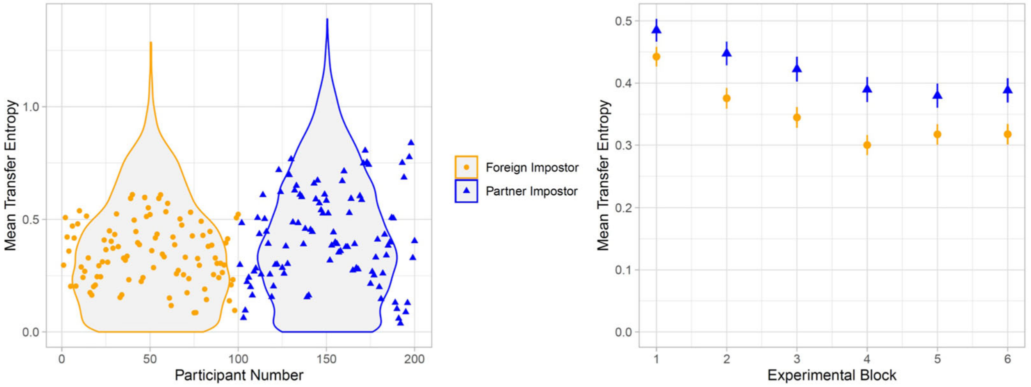

Fig. 7. Left: Distribution of transfer entropy by participants. Right: Development of the transfer entropy over experimental blocks. Error bars show the 95% confidence intervals.

3.3.4. Reciprocity

The mean transfer entropy (Fig. 7) was higher in the partner impostor condition ($M = 0.42$, $SD = 0.28$) than in the foreign impostor condition ($M = 0.35$, $SD = 0.24$), suggesting higher reciprocity. In the development over time, we surprisingly found a decrease instead of an increase in the partner impostor condition (Fig. 7), potentially due to the participants' issues with consistently communicating above chance.

3.4. Models

As preregistered, we tested our predictions in mixed-effects models (using *lme4*; Bates, Mächler, Bolker, & Walker, 2015) comparing the conditions and predicting performance or specific behaviors, respectively. The general analytical strategy was as follows: We always tried to control for the random effects of participants, nested in pairs, and bot seed, as well as for the fixed effects of trial number and partner identity. Where possible, we kept these effects maximal (Barr, Levy, Scheepers, & Tily, 2013), but we met convergence issues in some cases. When that happened, we tried to work around these issues by using a different optimizer, scaling the predictor variables, and simplifying the model (in this order). To simplify the model structure, we started by replacing the nesting for participants/pairs with unique participant ids, proceeded by removing the random effects correlation between slopes and intercepts, and finally removed random slopes (if nothing else helped). Subsequently, we will only mention issues that required a model simplification to converge. We assessed the significance of our results by model comparison to a simpler model that had the effect of interest removed, and compared the Akaike Information Criterion (AIC) of the relevant models. In line with Hooper, Coughlan, and Mullen (2008), we considered a ΔAIC greater than 2 as a meaningful difference between the models.

For prediction 1, we constructed a logistic regression model predicting accuracy by condition. In line with our prediction, the model comparison revealed that performance was better

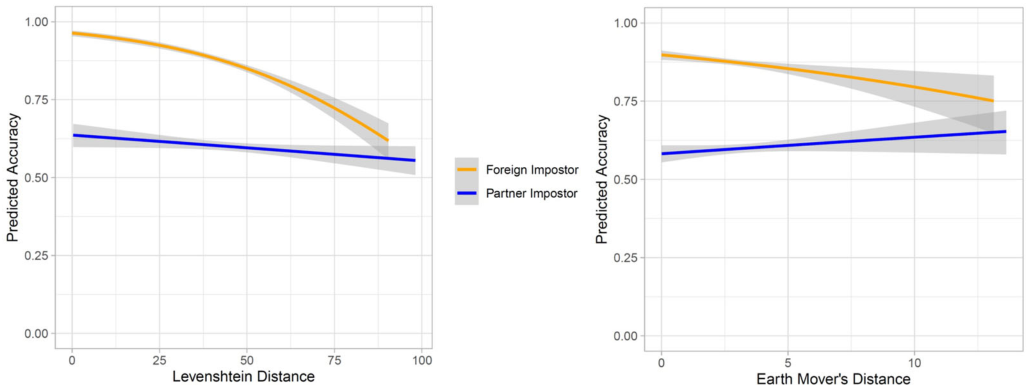

Fig. 8. Relation between conventionality and accuracy in both conditions. Lines represent the mean values smoothed by logistic regression; the shaded area is the 95% confidence interval around them. Note that higher distances mean lower conventionality. Left: Levenshtein distance measure. Right: Earth Mover's distance measure.

in the foreign impostor condition than in the partner impostor condition ($\beta = -1.92$, $SE = 0.16$, $\Delta\text{AIC} = 69$).

Research question 2 involved two predictions and two different measures to assess conventionality (the Levenshtein distance and the Earth Mover's distance). To test prediction 2.1, we built two logistic models predicting accuracy by conventionality, one for each of the distance measures. Here, the effect of interest was the interaction between condition and conventionality. In the model including the Levenshtein distance, we had to simplify the structure slightly for convergence by removing one slope–intercept correlation and the nested participant/pair structure. The results supported our prediction that participants in the foreign impostor condition would benefit more from conventional interaction (i.e., lower distance) than those in the partner impostor condition ($\beta = 0.55$, $SE = 0.10$, $\Delta\text{AIC} = 25$), even though it was not detrimental for the latter (see Fig. 8). The model including the Earth Mover's distance required only removing the nesting structure for convergence. Its interaction effect ($\beta = 0.30$, $SE = 0.08$, $\Delta\text{AIC} = 10$) showed that participants in the foreign impostor condition profited from conventionality and participants in the partner impostor condition suffered from it (Fig. 8), supporting prediction 2.1 even stronger.

As planned, we computed the two tests for prediction 2.2 only on the subset of data by pairs that performed above chance (as identified in the descriptive statistics). We then predicted the two distance measures by condition in separate linear models. For the Levenshtein distance, we did not find a meaningful difference between the two conditions ($\beta = 5.80$, $SE = 3.21$, $\Delta\text{AIC} = 0.9$), although descriptively the foreign impostor condition acted more conventionally. For the Earth Mover's distance, participants in the foreign impostor condition acted more conventionally than those in the partner impostor condition ($\beta = 0.78$, $SE = 0.28$, $\Delta\text{AIC} = 4.6$), supporting our prediction.

Research question 3 also involved two predictions, which we both addressed with our measure of transfer entropy. The test for prediction 3.1 was a logistic regression model predicting

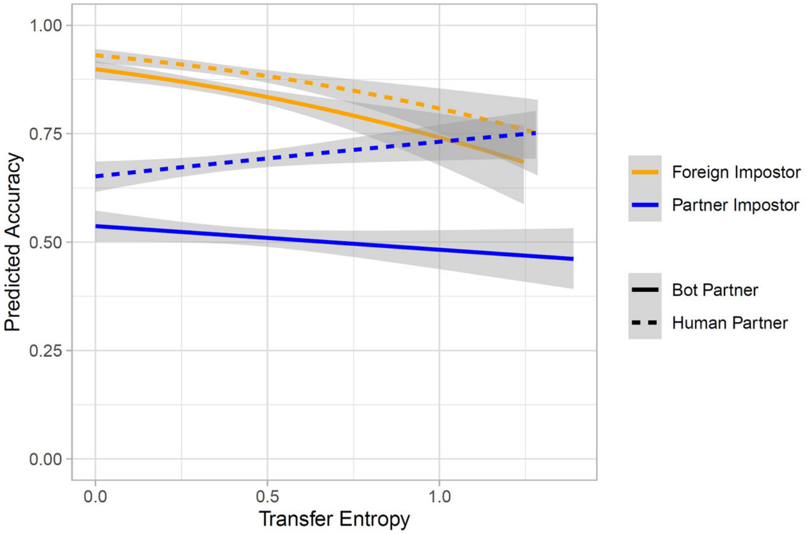

Fig. 9. Relation between reciprocity and accuracy in both conditions, split by partner identity. Lines represent the mean values smoothed by logistic regression; the shaded area is the 95% confidence interval around them. Note that high transfer entropy means high reciprocity.

accuracy by reciprocity (regardless of condition). The resulting effect was negative (opposite to our prediction), but negligible in size ($\beta = -0.08$, $SE = 0.04$, $\Delta\text{AIC} = 1.6$). However, we realized a problem for reciprocity that we did not account for in the preregistration: Its effects for accuracy depended on whether the current partner was the human or the bot. To benefit from reciprocity in the identification of the partner, the interactions with the partners have to differ in the first place. Therefore, we ran an exploratory follow-up analysis that also included partner identity, interacting with reciprocity. This more complex model required simplification by removing the slope–intercept correlations, and by removing the random slopes for the intercept on bot seed. It revealed that reciprocal behavior was beneficial when interacting with the human, but detrimental when interacting with the bot ($\beta = 0.17$, $SE = 0.07$, $\Delta\text{AIC} = 3.4$). These results suggest that participants might have treated reciprocity as a signal for “humanness”. Further exploration by condition showed that the effect was driven solely by the partner impostor condition (Fig. 9). Therefore, we can summarize that reciprocity was indeed helpful for successful communication when establishing it with the human partner (but not the bot), and only in the partner impostor condition.

Prediction 3.2 was also tested on the subset of participant pairs performing above chance. We predicted the transfer entropy by condition in a linear model. We found that participants in the partner impostor condition were acting more reciprocally than participants in the foreign impostor condition ($\beta = 0.10$, $SE = 0.04$, $\Delta\text{AIC} = 3$), supporting our prediction.

Finally, we constructed a model combining condition, conventionality, and reciprocity as predictors of accuracy to investigate their respective importance in a single step. Because of the complexity of this model, we needed to remove slope–intercept correlations, the nested group/participant structure, and reduce the model to the two most important random slopes (reciprocity and conventionality) for it to converge. We did not test anything in particular in this model, since its purpose was to explore, but found that the previous results held up even when put together at once: Participants were overall worse in the partner impostor condition, conventionality was helpful for success only in the foreign impostor condition, and reciprocity was increasing success when it happened with the human partner in the partner impostor condition.

4. Discussion

Taken together, our results show that machine impostors can become more deceptive and impede the successful formation of novel conventions by imitating past interactions—the main outcome of our study. We conclude this from the combined findings from research questions 1 and 2: As evidenced by the consistently higher performance in the foreign impostor condition, access to the human partner's past behaviors made bot impostors appear more deceptive. The average increase in performance in the foreign impostor condition, compared to the relative stagnation in the partner impostor condition, already suggests that participants successfully created conventional solutions in the former but not in the latter condition; as does the direct increase in conventionality (as measured) over time. This suggestion is corroborated by the statistical modeling results for conventionality in predictions 2.1 and 2.2: For both our measures, more conventional behaviors predicted better performance in the foreign impostor condition, compared to the partner impostor condition. For the Earth Mover's distance, conventional behavior in the partner impostor condition was even detrimental. It makes sense then that, judging by this measure, successful participants in the foreign impostor condition also acted more conventionally than their counterparts in the partner impostor condition, as shown by the model for prediction 2.2. In sum, mimicking a participant's partner instead of a foreign player made bot impostors more deceptive and conventionality a less useful option for communicative success.

The results for reciprocity were more complex than expected, yet complementary to this. Against our expectations, more reciprocal behavior did not lead to improved performance overall. However, the more complex exploratory model revealed that reciprocity did help when it happened with the human partner (as opposed to the bot), but this was only the case in the partner impostor condition. Adding to this is the result for prediction 3.2, showing that above-chance participants in the partner impostor condition acted more reciprocally than their counterparts in the foreign impostor condition (as predicted). Putting these results together, we can state that reciprocity was only useful in the partner impostor condition (if it distinguished humans from bots), and that the relatively few successful pairs in this condition managed to rely on it, at least to some degree.

These results suggest that conventionality and reciprocity are adaptive under the right conditions. Specifically, conventionality helped establish communication in the foreign impostor condition and reciprocity helped in the partner impostor condition. The results thus show empirical support for these two potential routes to the successful emergence of communication and shine light on their underlying mechanisms. Conventionality being "blocked" from successful usage in the partner impostor condition was a key result intended by our manipulation. Since the impostor bot in this condition had the ability to catch up to participants' latest behaviors and copy them a few rounds later, it posed a constant challenge to flexibly update communication or risk deception. This highlights the reliance on precedence for conventional routes to communication. In contrast to the conventionality results, we did not expect reciprocity to be only viable in the partner impostor condition; yet, we did expect that this condition would be more reliant on reciprocity. Here, participant pairs could overcome bot impostors by demonstrating live interaction (as opposed to the unreactive bot) through reacting to one another's movements. Potentially, the easy availability to communicate successfully through conventional methods might have made this option obsolete or outright detrimental in the foreign impostor condition, since it would distract from the simplest strategy: Repeat what was proven successful already.

It is important to note the potentially amplifying role of feedback in this regard, a much discussed topic in experimental semiotics (e.g., Garrod et al., 2007; Healey et al., 2007; Müller et al., 2019). Participants received perfect information about their own and their partner's performance after each trial in the experiment. This was a deliberate design choice to encourage collaboration and motivate players, but also necessary since piloting revealed that the task was very hard for the average pair in either condition. However, it might have also emphasized conventional routes to success by reinforcing whatever behavior was successful before. A possible benefit of this reinforcement could be that it supported the relatively arbitrary conventions we aimed for in this study, but it could also explain why reciprocity only really mattered once conventionality was impossible. In any case, future research is well placed to use a different version of this paradigm and systematically investigate the role of feedback for the emergence of conventional strategies.

An interesting point regarding the interpretation of the study is the nature of the impostor bots. These bots were implemented by faithfully replaying previously recorded human behavior. In this sense, there is only a very limited algorithm guiding their behavior, outside of what source to pick the replay from. However, we believe that this is not a problem for the ecological validity of the study: First, participants did not have mechanistic insight into bots' behaviors, with no prior instruction on what to expect from the bots. They sometimes suspected that the bot was copying human behavior; if so, they often believed it was copying their own moves (an interesting point by itself), and this also happened in the foreign impostor condition, where bots were not even drawing from their own dyad. Second, real-world bots are also increasingly learning from human data rather than being rule-based. This can be likened to the situation in our study, even if they are informed by data in a more complex way and not completely copying it.

Third, authentic human behavior is arguably the upper bound of performance for a machine in this task and should be particularly hard to differentiate for participants. In this way, we circumvented the potential issue that the bot might not be good enough at deception, and avoided a black box of bot behaviors we do not understand. Yet, we deliberately kept the limitation that the bots could not react to the current participant's behavior, opening up space for reciprocal solutions to the task. This is reproducing the lack of pragmatic capabilities preventing bots from passing the Turing test in natural language (Jacquet et al., 2021). An exciting avenue for future research would be to deviate from this constraint and attempt to construct bots that can mimic the reciprocal strategies. In our view, it should be possible to build this “pragmatic AI” in this specific minimal task. Most likely, it would turn out to be unbeatable by humans, constituting a small step toward bots centered on interaction, not language in a narrow sense.

Regarding the relevance of our study for real-life bots that we alluded to in the introduction, the current results suggest potential dangers in a scenario where bots manage to integrate personal information into deception attempts. Personalized recommender systems are already a reality, so it is imaginable that fraudulent bots could also take advantage of the information they can access, for example, within a social network. A concrete example would be a bot mimicking a typical conversation with an online acquaintance to lure victims onto a fraudulent website or to ask for money, for instance, via hacked or imitated social network profiles. Our results would suggest that such an attempt could be more dangerous than the typical spam and fake profiles we encounter online. In real life, humans navigate complex social and communicative environments that have the potential to protect us from deception through simple bots, but with the lower complexity in virtual interactions, this risk increases. Even though our task reduced interaction to its minimal foundations and did not involve natural language, it is a relevant concern precisely because we studied the basic abilities human communication is built on. On a more positive note, our results also show how reciprocal communication can be a solution to detect unreactive artificial agents reliably.

In the bigger picture, our study is an example of how a Turing test setup can be a fruitful approach to assess the basic principles of human communication. It has been argued that the Turing test should be more widely adopted to study reasoning and pragmatics (Jacquet et al., 2021). We expand on this idea by removing the requirement of natural language, opening up space to investigate the emergence of communication, based on ideas from pragmatics. Possible extensions include scaling up the complexity of the task and increasing the sophistication of the bots to study higher-level phenomena. The key strength and novelty we see in Turing test adaptations is that bots are controlled by the experimenter, and can thus be manipulated to test hypotheses about how human interaction and coordination can function. In contrast to confederates, these automated agents can deliver interactive behavior reliably in the same way. Simultaneously, we also learn about which characteristics will make interactive bots more convincing and human-like, or at least gain new insights into the mechanisms of machine-human interaction. Ultimately, our study is uniting ideas from language evolution, social cognition, and pragmatics within a new experimental task.

5. Conclusion

How can humans successfully establish communication when faced with imitative machine impostors that attempt to deceive them? We set out to investigate the role of conventions and reciprocal interactions for emergent communication in a minimal Turing test setup. Participants attempted to find out about their partner’s identity only through their virtual movements. Our main findings are that bot impostors were more deceptive when they accessed the personal interaction history of a pair, and that they impeded the successful formation of conventions in this situation. Yet, under this condition, we found that reciprocity was a useful strategy. These results shed some light on human routes to communication even in minimal environments, and imply an increased risk through bots online, should those make use of these communicative mechanisms.

Acknowledgments

We would like to thank Deniz Gezerli for his help with the data preprocessing and creating the gifs in the Supplementary Material, and Kira von Kleist for the data collection on Prolific. An additional thanks goes to Jonas Mücke for his help with monitoring and stabilizing the experiment on the server. Last, we want to thank all pilot testers who kindly volunteered and gave feedback on early builds of the game: Sophie-Marie Rostalski, Barbara Pavlek Löbl & Mathias Löbl, Claudia & Stefan Geßlein, Artjoms Šeļa, Jason Bassett, Mengchen Dong, Maria Brackin (and Ada!), Stefanie Eisenmann, Deniz Gezerli, and Kira von Kleist.

Open access funding enabled and organized by Projekt DEAL.

Open Research Badges

This article has earned Open Data and PreRegistered badges. Data are available at https://osf.io/z4vtj and the preregistration is available at https://osf.io/txgh7.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171. https://doi.org/10.3758/s13428-020-01535-9

Al-Rousan, S., Abuhussein, A., Alsubaei, F., Kahveci, O., Farra, H., & Shiva, S. (2020). Social-guard: Detecting scammers in online dating. In 2020 IEEE International Conference on Electro Information Technology (EIT) (pp. 416–422). https://doi.org/10.1109/EIT48999.2020.9208268

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, *27*(1), 32–47. https://doi.org/10.1016/j.newideapsych.2007.12.002

Auvray, M., & Rohde, M. (2012). Perceptual crossing: The simplest online paradigm. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00181

Barone, P., Bedia, M. G., & Gomila, A. (2020). A minimal Turing test: Reciprocal sensorimotor contingencies for interaction detection. *Frontiers in Human Neuroscience*, *14*(102). https://doi.org/10.3389/fnhum.2020.00102

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedia, M. G., Aguilera, M., Gómez, T., Larrode, D. G., & Seron, F. (2014). Quantifying long-range correlations and 1/f patterns in a minimal experiment of social interaction. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01281

Behrendt, S., Dimpfl, T., Peter, F. J., & Zimmermann, D. J. (2019). RTransferEntropy—Quantifying information flow between different time series using effective transfer entropy. *SoftwareX*, *10*, 100265. https://doi.org/10.1016/j.softx.2019.100265

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE*, *7*(8), e43807. https://doi.org/10.1371/journal.pone.0043807

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994. https://doi.org/10.1073/pnas.1418838112

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41. https://doi.org/10.1080/01690968708406350

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLoS ONE*, *10*(9), e0136100. https://doi.org/10.1371/journal.pone.0136100

European Commission (2021). Regulation COM/2021/206 final: Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. Retrieved from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

Froese, T., Iizuka, H., & Ikegami, T. (2015). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports*, *4*(1), 3672. https://doi.org/10.1038/srep03672

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767. https://doi.org/10.1207/s15516709cog0000_34

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00011

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215. https://doi.org/10.1016/0010-0277(94)90048-5

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987. https://doi.org/10.1080/03640210701703659

Healey, P. G. (2008). Interactive misalignment: The role of repair in the development of group sub-languages. In R. Cooper & R. Kempson (Eds.), *Language in flux* (pp. 13–39). Palgrave Macmillan.

Healey, P. G., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, *31*(2), 285–309. https://doi.org/10.1080/15326900701221363

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Iizuka, H., Ando, H., & Maeda, T. (2009). The anticipation of human behavior using "parasitic humanoid." In J. Jacko (Ed.), *International Conference on Human–Computer Interaction* (pp. 284–293). Springer.

Iizuka, H., Marocco, D., Ando, H., & Maeda, T. (2013). Experimental study on co-evolution of categorical perception and communication systems in humans. *Psychological Research*, *77*(1), 53–63. https://doi.org/10.1007/s00426-012-0420-5

Jacquet, B., & Baratgin, J. (2021). Mind-reading chatbots: We are not there yet. In T. Ahram, R. Taiar, K. Langlois, & A. Choplin (Eds.), *Human interaction, emerging technologies and future applications III* (Vol. 1253, pp. 266–271). Springer International Publishing. https://doi.org/10.1007/978-3-030-55307-4_40

Jacquet, B., Jamet, F., & Baratgin, J. (2021). On the pragmatics of the Turing test. In *2021 International Conference on Information and Digital Technologies (IDT)* (pp. 123–130). https://doi.org/10.1109/IDT52577.2021.9497570

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100. https://doi.org/10.1145/1290958.1290968

Kopp, S., & Krämer, N. (2021). Revisiting human–agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, *12*, 580955. https://doi.org/10.3389/fpsyg.2021.580955

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, 113–114. https://doi.org/10.3758/BF03342817

Lenay, C., & Stewart, J. (2012). Minimalist approach to perceptual interactions. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00098

Lenay, C., Stewart, J., Rohde, M., & Amar, A. A. (2011). "You never fail to surprise me": The hallmark of the Other: Experimental study and simulations of perceptual crossing. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *12*(3), 373–396. https://doi.org/10.1075/is.12.3.01len

Lewis, D. (1969). *Convention*. Harvard University Press.

Makowski, D. (2018). The *psycho* package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, *3*(22), 470. https://doi.org/10.21105/joss.00470

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213. https://doi.org/10.1016/S0749-596X(03)00028-7

Müller, T. F., Winters, J., & Morin, O. (2019). The influence of shared visual context on the successful emergence of conventions in a referential communication task. *Cognitive Science*, *43*(9), e12783. https://doi.org/10.1111/cogs.12783

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

Pinar Saygin, A., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, *10*(4), 463–518. https://doi.org/10.1023/A:1011288000451

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

dos Santos, M., Matias Rodrigues, J. F., Wedekind, C., & Rankin, D. J. (2012). The establishment of communication systems depends on the scale of competition. *Evolution and Human Behavior*, *33*(3), 232–240. https://doi.org/10.1016/j.evolhumbehav.2011.09.007

Schelling, T. C. (1960). *The strategy of conflict*. MIT Press.

Scott-Phillips, T. C. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Palgrave Macmillan.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417. https://doi.org/10.1016/j.tics.2010.06.006

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233. https://doi.org/10.1016/j.cognition.2009.08.009

Seymour, J., & Tully, P. (2016). Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. Retrieved from https://tutorial.evogtechteam.com/wp-content/uploads/2017/03/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, *39*(1), 212–226. https://doi.org/10.1111/cogs.12150

Sperber, D., & Wilson, D. (1994). Relevance: Communication and cognition *(Reprint)*. Blackwell.

Tamariz, M. (2017). Experimental studies on the cultural evolution of language. *Annual Review of Linguistics*, *3*(1), 389–407. https://doi.org/10.1146/annurev-linguistics-011516-033807

Tinits, P., Nölle, J., & Hartmann, S. (2017). Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution*, *2*(2), 148–159. https://doi.org/10.1093/jole/lzx011

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Urbanek, S., & Rubner, Y. (2012). emdist: Earth mover's distance. Retrieved from https://CRAN.R-project.org/package=emdist

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449. https://doi.org/10.1017/langcog.2014.35

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30. https://doi.org/10.1016/j.cognition.2018.03.002

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, *154*, 102–117. https://doi.org/10.1016/j.cognition.2016.05.011

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.